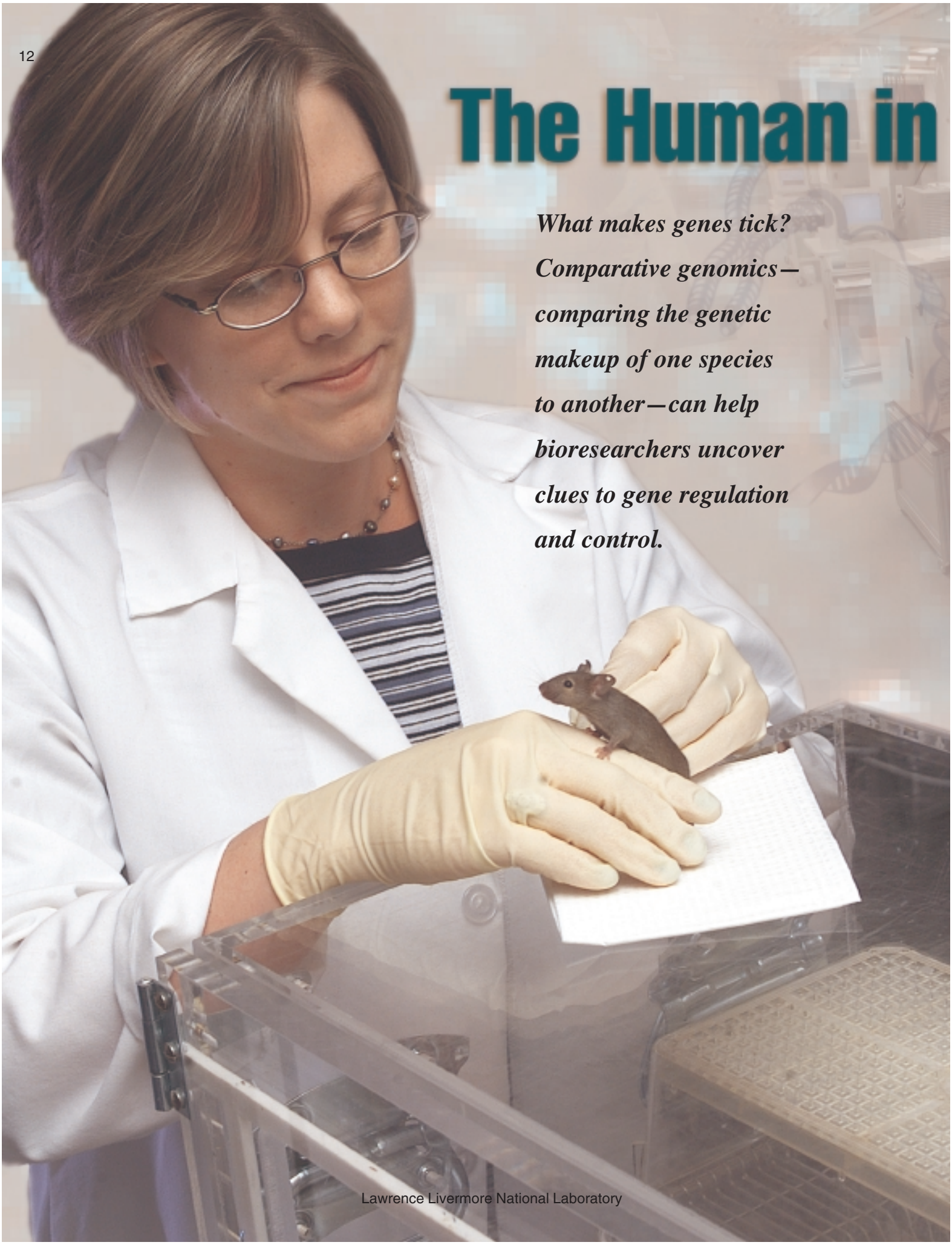
The Human in

What makes genes tick? Comparative genomics—comparing the genetic makeup of one species to another—can help bioresearchers uncover clues to gene regulation and control.



the Mouse Mirror

IN the excitement over the completed draft sequence of the human genome—certainly a grand accomplishment—it's easy to forget that this is just the prologue. Much about the genome remains a mystery. Which parts of it are actual genes? What do individual genes do, and how do they do it? (See the box on p. 17.) A small, four-footed mammal—the mouse—is helping to answer these questions. By comparing the human and mouse genomes piece by piece, bioresearchers such as Lawrence Livermore's Lisa Stubbs are uncovering clues to genomic mysteries.

After the draft sequence for the human genome was completed last June (see the box on p. 18), the Department of Energy's Joint Genome Institute (JGI) turned to sequencing pieces of mouse DNA that correspond to human chromosome 19. "We focused on this particular human chromosome because the Laboratory has created an extremely thorough gene map for it over many years of research," says Stubbs. "The sequence is not finished yet, but its working draft is easier to read than the draft sequence of many other human chromosomes. Because of the careful way the map was constructed, we know the sizes of the gaps in the chromosome and the way the pieces fit together."

Since last October, when the mouse sequencing was completed, Stubbs and her team have been analyzing the mouse and human DNA sequences, examining both similarities and differences to discover what the sequences reveal about our genes and our genetic evolution.

Comparing the two sets helps the scientists track down genes—which are not always easy to spot—and provides information about the nongene portions of DNA that make up nearly 99 percent of our genome. Beyond that, having an understanding of why and how mouse and human genomes are different provides critical information to the bioscience and medical research communities. Stubbs explains, "If we're going to use the mouse as a model for the human, which everybody is doing, we'd better know how the two species differ and try to answer questions such as: How often do human and mouse contain the same genes? How similar are the genes? Are there exceptions to the rule of similarity? We must know these things on a gene-by-gene basis because while some genes are very similar, others are not. Knowing all this will help us understand whether it's right to use mice for drug testing and as disease or drug models. And if it's not right, why not? Even the 'why nots' reveal something about the human gene and how it works."

Junk, Shattered Genes, and a Twist

Two intriguing elements of the human genome came to light as a direct result of this comparative genomics: the different sizes of some related human and mouse regions and the composition of "junk" between the genes. Two pieces of related DNA for mouse and human show more or less the same genes in more or less the same order. But when Stubbs and her team spread out the two sequences and laid them side by side—the first time this has been done on a chromosome-wide scale—they discovered that many human regions are significantly larger and less compact than the mouse regions. So what's the filler in the human sequence? Scientists refer to it as junk, but not just any junk.

"For instance," Stubbs says, "there is a particular kind of junk sequence called the Alu sequence. It's a repetitive DNA sequence that, in the human, has made lots and lots of copies of itself and has

infected our DNA to a much greater extent than anything we see in mouse. It's just one of many DNA junk elements that make copies of themselves and litter the human genome in the millions."

Repetitive sequences like Alus are essentially DNA parasites. Their duplication generally does not appear to have serious functional consequences, although Alu copies that get inserted into genes have been shown to cause human disease. Stubbs notes that this sort of litter is also seen in mouse DNA. However, the Alu sequence invasion shows up more recently in the evolution of DNA and appears to have occurred more dramatically in the primate than the rodent lineages. Because mouse and human evolution haven't been separated all that many years, the difference in overall size and

Bioinformaticist Paramvir Dehal sits in front of a computer screen showing results of a genome sequence analysis tool he developed (see the box on p. 18 for more details). Part of the results show a comparison of chromosome 19 from the human genome and a related piece from the mouse genome sequenced by Livermore bioresearchers. Only about 7 percent of the two genomes appear related, and within that 7 percent, only about 60 percent are genes. Comparative sequencing helps researchers zoom in on the important, conserved (unchanged) sections of genomes.

amount of junk is remarkable. "This is something we wouldn't have seen if we hadn't been able to lay out the pieces of sequence and compare them," she said. Why junk sequences happen and what they mean remain to be seen.

When the mouse and human sequences are compared, other broad similarities and differences quickly become apparent. Of the small percentage of the parts that make up genes, about 85 percent appear to be the same in sequence for both species. In addition, both mouse and human have basically the same number of genes generating more or less the same kinds of proteins. However, the genes lying on human chromosome 19 show up on several different mouse chromosomes. It's as if someone shattered the human chromosomes and rearranged blocks of 20 to 200 genes into different orders to produce the mouse genome.

"This sort of rearrangement happens in evolution," says Stubbs, "but when we look at the genomes of other mammals that are just as far removed in evolution from the human as the mouse—the cat, dog, or cow—their chromosomes are much more similar to ours than the chromosomes from the rodent family. So what drives the breakup of mouse chromosomes? There are several theories, most concerning the short generation time and breeding habits of rodents, but what it comes down to is, we don't know yet."

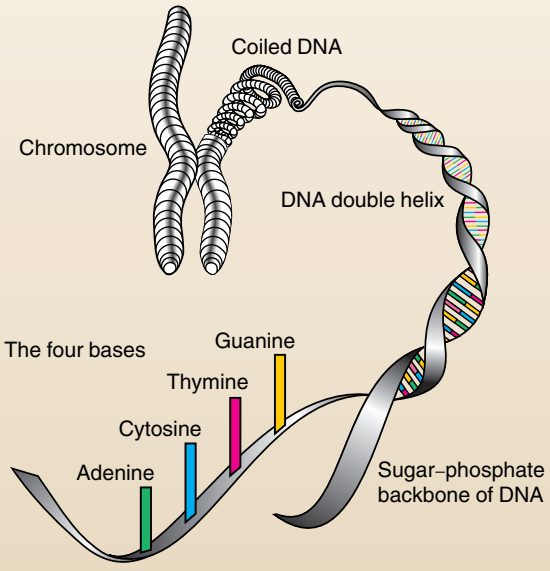
In another interesting twist, when mouse and human genes were compared, quite a number of human-specific and mouse-specific genes were found. These species-specific genes are altogether a small fraction of our 30,000 genes, but still a significant number, probably several hundred genome-wide. "We—and nearly everyone else—expected to find a nearly one-to-one correspondence between mouse and human genes," says Stubbs. "The species-specific genes are of several different types, but the largest number of them appear to make or express regulatory proteins that do the actual business of turning genes on and off."

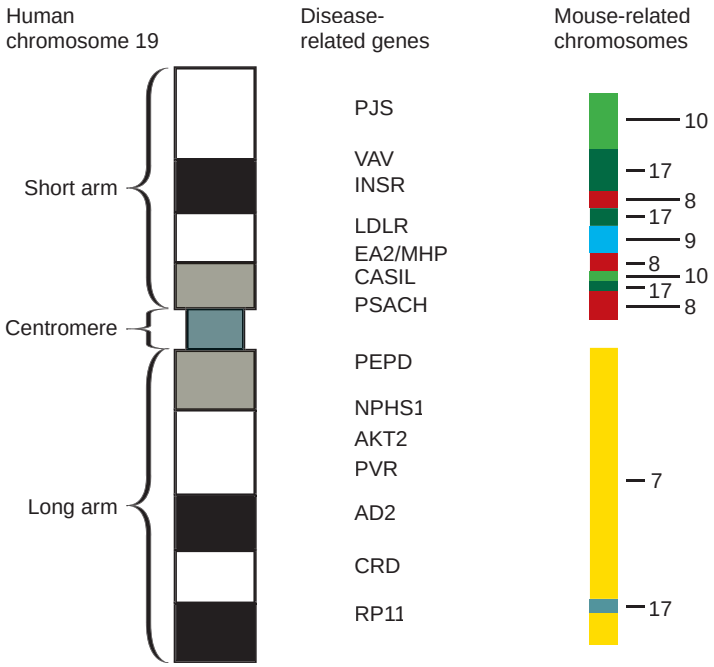
These proteins, continues Stubbs, are probably not critical, meaning that

Genome Basics

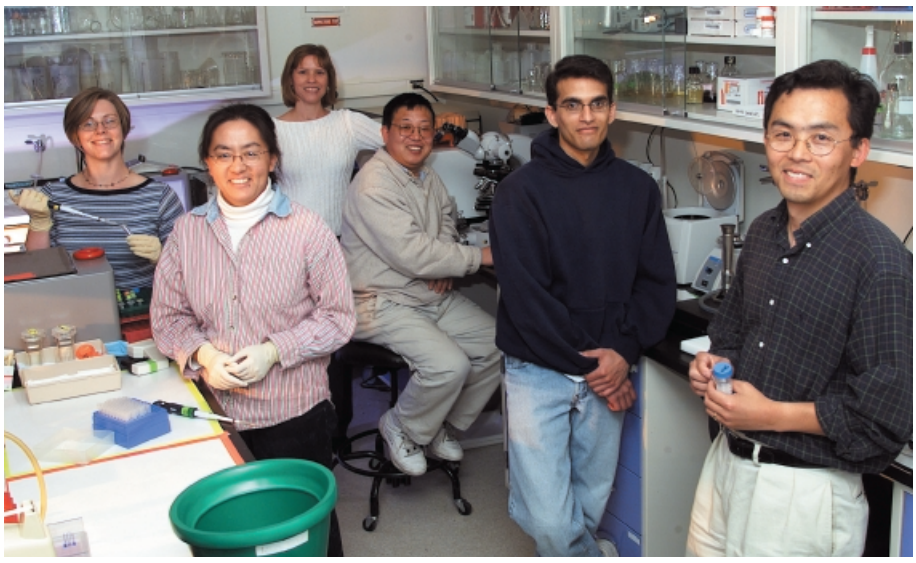
Each human cell contains 23 pairs of chromosomes in its nucleus. Each chromosome contains two tightly coiled strands of DNA (deoxyribonucleic acid), with each DNA strand composed of "base pairs" of chemical bases, normally abbreviated A, C, T, and G (for adenine, cytosine, thymine, and guanine). Scientists estimate that about 3 billion base characters comprise the human genome, with about 1.5 percent of those characters forming genes. Genes are special stretches of DNA that carry a code for making proteins, which are critical to helping our cells function. The process for making proteins is exact. Each cell contains complex proteins called transcription machinery. When it is time for a protein to be made, these machines go into the nucleus, find the control sequences that signal a particular gene to start, and bind to them. The transcription machinery then makes a mirror copy, or transcript, of the gene's sequence, as indicated by the control. The transcript, referred to as RNA (ribonucleic acid), then moves out of the nucleus and into the cell's cytoplasm where it encounters another biological machine, the ribosome. The ribosome, using the RNA as a set of instructions, assembles a protein from amino acids.

One way scientists identify genes is to capture RNA sequences in the cytoplasm and analyze them to determine which DNA sequences correspond to which RNA sequences. These captured RNA sequences are called complementary DNA (cDNA) sequences, and numerous collections of cDNA sequence snippets, called expressed sequence tags, are available in public databases. "A cDNA is a copy of the gene," explains Livermore bioresearcher Lisa Stubbs. "Bioscientists have found ways to take RNA out of the cells, 'reverse transcribe' them into cDNA copies, clone them into bacteria, and sequence them. From the reverse transcription, we get a snapshot of the sequences in a particular cell that are being turned on and turned into proteins at a particular time."





The human and mouse genomes are both similar and different. The long arm of human chromosome 19 has a close counterpart in mouse chromosome 7—the human and mouse versions of the same genes (see middle column) are found in them in roughly the same order. However, genes in human chromosome 19's short arm correspond to mouse versions that are located in many different mouse chromosomes, as indicated by the colored bars to the right, labeled by chromosome number.



Some of the members of the mouse genomics group are, from left, Laura Chittenden, Xiaojia Ren, Lisa Stubbs (team leader), Xiaochen Lu, Paramvir Dehal, and Joomyeong Kim.

gaining or losing them will probably not result in disaster to the organism. Instead, they probably are involved in fine-tuning traits. "These species-specific genes are very likely to be a major source of subtle diversity and keys to the subtle differences in gene expression between species," she says. Although the effects of changing a single gene are probably small, the combined effects of hundreds of changes are likely to be significant.

What Makes Humans Human?

Whether a gene resides on chromosome 2 or 20 usually does not affect its function. (The main exceptions to this rule are the genes on the sex-linked chromosomes X and Y.) That being said, scientists have to question why, with mice and humans having almost identical sets of 30,000 genes, they aren't more alike. Part of the answer is that a 15 percent difference in the sequence of a gene can change its function dramatically. For example, many human genes that cause disease differ from their normal counterparts by a single nucleotide. For most genes, this nucleotide change would constitute less than a 0.1-percent sequence change, but the result is a devastating functional difference.

Take the PEG3 gene, which is shared by mouse and human. It plays an important role in embryonic mouse development and an even more important role in mouse maternal behavior. Research shows that when the PEG3 gene is removed from mice, the mothers ignore their young to the point that their babies die. A similar protein is expressed in the human brain, says Stubbs, so the maternal caring function is probably conserved—unchanged during evolution—to some extent. "However, the levels of expression differ—the protein is expressed like gangbusters in the mouse brain, not so highly in the human. Even more intriguing, it's highly expressed in

human ovaries and placentas, but not at all in mouse ovaries. It seems likely that this gene has taken on a role in humans that it isn't playing in mice."

Stubbs notes that many similar mouse and human genes have differing behavior: activated in one kind of tissue in mouse but not in human, or perhaps appearing in the same tissue in both, but at different times or with different intensities. "In other words, the same genes are not necessarily regulated or controlled in the same way in both species. The dissimilarities may be part of the answer as to why mice are mice and humans are human."

So what controls the on–off switch in genes and the timing of gene expression? Here again, rodents provide some clues. When researchers compare human and mouse sequences, they find small sections that are similar between the species but are not genes or junk such as Alus or other identifiable repetitive elements. Stubbs explains, "We can look at a piece of sequence and see that it isn't making part of a protein—so it isn't part of a gene. These mystery pieces, like genes, stand out as conserved DNA against a nearly 95-percent background of totally dissimilar sequence and are good candidates for a control sequence." Researchers know little about these types of sequences except that they are extremely important, hard to detect, and have been conserved because their sequence is linked to function. Many researchers are beginning to explore control sequences now that there is a way to find them through their conservation (because human and mouse genome sequences are known). Gene regulation, Stubbs says, is turning out be one of the most exciting areas of current research in the field.

Looking Section by Section

Learning more about control sequences and other regulatory elements in gene expression is one of the next genomic frontiers. One technique used by the biomedical research community is tissue-section analysis, which is related to a standard hospital biopsy technique. The technique involves slicing 10-micrometer-thick sections of tissue (about the thickness of a single cell). It permits single cells to be viewed in their native context using microscopy and standard pathological techniques.

Adopting this technique, Stubbs and her team place thin slices of fetal or adult mouse tissue on a slide and add a gene probe, which is a specific gene sequence to which a fluorescent dye has been added. The probe binds to the unique RNA sequence produced by the gene under study. (The RNA—ribonucleic acid—is a mirror image of the DNA sequence of a gene and an intermediate in the process of protein coding.) When the tissue is observed under a microscope, the fluorescent probe can be seen binding to and highlighting the cells in which the particular RNA has been expressed. This technique of highlighting cells is called in situ hybridization.

Because a mouse fetus in even the latest stages of development is only about l centimeter long, its entirety can fit on a slide to give researchers a whole-body picture of where a particular gene is expressed. Stubbs explains, "Our pathologist Xiaochen Lu can look at a single specimen and tell us what cells are activated and what the purpose of those cells is. So if that gene is turned

Laying Out the Human Genome

In February, the International Human Genome Sequencing Consortium—of which the Department of Energy's Joint Genome Institute (JGI) is a part—and the commercial company Celera simultaneously published papers in the scientific journals *Nature* and *Science* describing the draft sequencing of the human genome. The initial analysis of this draft sequence held a number of surprises. All in all, there appears to be only about 30,000 genes, equaling about 1 to 1.5 percent of the sequence. In other words, in the nearly 2-meter-long strand of DNA that appears in each and every cell of our bodies, about 15 centimeters of it contain genes. The number of genes is about a third to a half of what most scientists had believed would be the case. As Trevor Hawkins, JGI director, noted, "It puts us humans at something like about twice as many genes as your average fruit fly, which, I think, is quite a humbling thought."

Most of the leftover 99 percent of our DNA appears to be junk, or at least DNA whose functions remain unknown. Littered in the junk are long sequences similar to those found in viruses and bacteria. These sequences appear to have taken up residence in the genome as far back as 700 million years ago, when life was composed of a single cell. "These sequences clearly have the structure of viral DNA," explains bioresearcher Lisa Stubbs, "but they've lost the ability to turn into a virus particle."

The International Human Genome Sequencing Consortium includes 20 groups from the United States, the United Kingdom, Japan, France, Germany, and China. Among those groups is the JGI, a virtual institute that integrates the sequencing activities of the human genome centers at Lawrence Livermore, Lawrence Berkeley, and Los Alamos national laboratories. For more information about the initial analysis and sequencing of the human genome by the International Human Genome Sequencing Consortium, see www.nature.com/genomics/human/.

Tools of the Comparative Trade

Organizations such as the Joint Genome Institute (JGI) are extremely proficient in sequencing DNA, turning a task that used to be done painstakingly by hand a quarter century ago into an industrial procedure. However, analysis of that sequence—particularly comparing the sequence of two species—remains in the domain of human interpretation. Livermore bioresearcher Lisa Stubbs notes that there are computer programs to help scientists align the DNA sections of interest and to visualize similarities and differences. Computer algorithms can also identify a piece of DNA as a probable gene. But these tools are right only about 60 to 70 percent of the time and require human confirmation. Among the few computer tools available to help scientists visualize the differences and similarities between the sequence of two species are percent identity plots (PIPs) and dot plots.

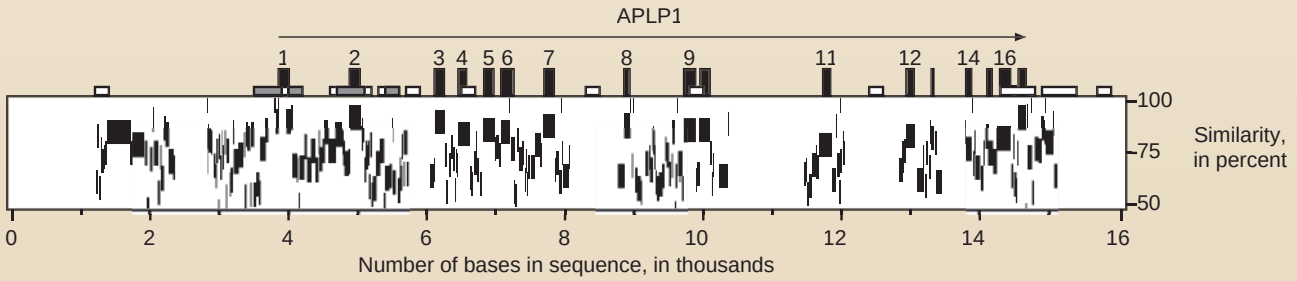
The PIP, developed by Webb Miller at Pennsylvania State University, is often used to find genes and regulatory elements. A scientist sends a file representing the bases of a piece of human sequence to the computer, followed by the piece of mouse DNA that corresponds to it. The program plots out the matches within the sections, marking matches with a dot and plotting them on a scale showing how similar the two sections are. Scientists can look along a stretch of DNA and quickly see that one piece is conserved—that is, hasn't changed during evolution—and then there's another little stretch of DNA that is somewhat less conserved and so on. The PIP program allows them to see how far apart those matches are. The program also can plot out positions of repetitive elements and find stretches of DNA that are rich in C and G bases. "We call these CpG islands," Stubbs explains, "Often, for some reason we don't yet understand, these islands are associated with control sequences. If you find an area rich in CpGs in both human and mouse, fairly close to a gene, it's a good candidate for a control sequence."

Dot plots are another tool that can be used to plot mouse DNA against the related piece of human DNA. In dot plots, the order of matching sequences of human and mouse DNA can be compared. Where the two aligned sequences match, a little mark is added to the graph. "This helps us see how the genomes align, where the similarities and differences in structure occur. For example, dot plots help us pinpoint the spots where the mouse chromosome has shattered, and half of it matches chromosome 19 and half matches another human chromosome," Stubbs says. "It helps us find those breaking points."

Stubbs notes that tools such as PIPs and dot plots are slow and are better suited for looking at small pieces of sequence. At the JGI, Paramvir Dehal, a bioinformaticist and Ph.D. candidate in the Department of Genetics at the University of California at Davis, is working with Stubbs, computer scientist Art Kobayashi, and others to develop tools for examining and analyzing larger pieces of sequence. The tools they develop will be specifically designed as aids for comparative genomics. One sequence analysis tool being developed by Dehal uses a color code to show areas of similarity among various types of sequence, whether human, mouse, *Drosophila* (fruit fly), flatworm, yeast, or expressed sequence tags. A yellow bar along the chromosome map means the human DNA at that site has similarity to DNA from another species or to a recognized, previously studied human gene. Clicking on the bar brings up another screen that shows details of the sequence matches at that site and the degree of similarity between the matches, which is indicated by its colors. Red means an almost identical match; pink indicates a related sequence, but not a perfect match; and green or blue indicates that the matching sequence has few similarities to the human DNA.

Scientists can use this tool to find out which areas of the sequence are conserved among species. Areas of conservation usually indicate an important function, whether the area is a gene, regulatory sequence, or something else. "A pink match to *Drosophila* is truly significant because flies and humans are so far removed from each other in evolution. The likelihood is high that such a highly conserved piece of DNA is coding for a protein," Stubbs notes.

The tool is also handy for hunting down regulatory or control sequences. A piece of human sequence is a good candidate for a regulatory sequence if it matches mouse DNA, but not a cDNA sequence, and does not appear to be encoding a protein. Experiments must be done to verify the function of a conserved sequence because scientists presently cannot really predict a piece of DNA's function just by looking at its sequence. However, conservation does tell them which sequences are important and points them to the 1 to 5 percent of the genome they should focus on, which is an important first step.

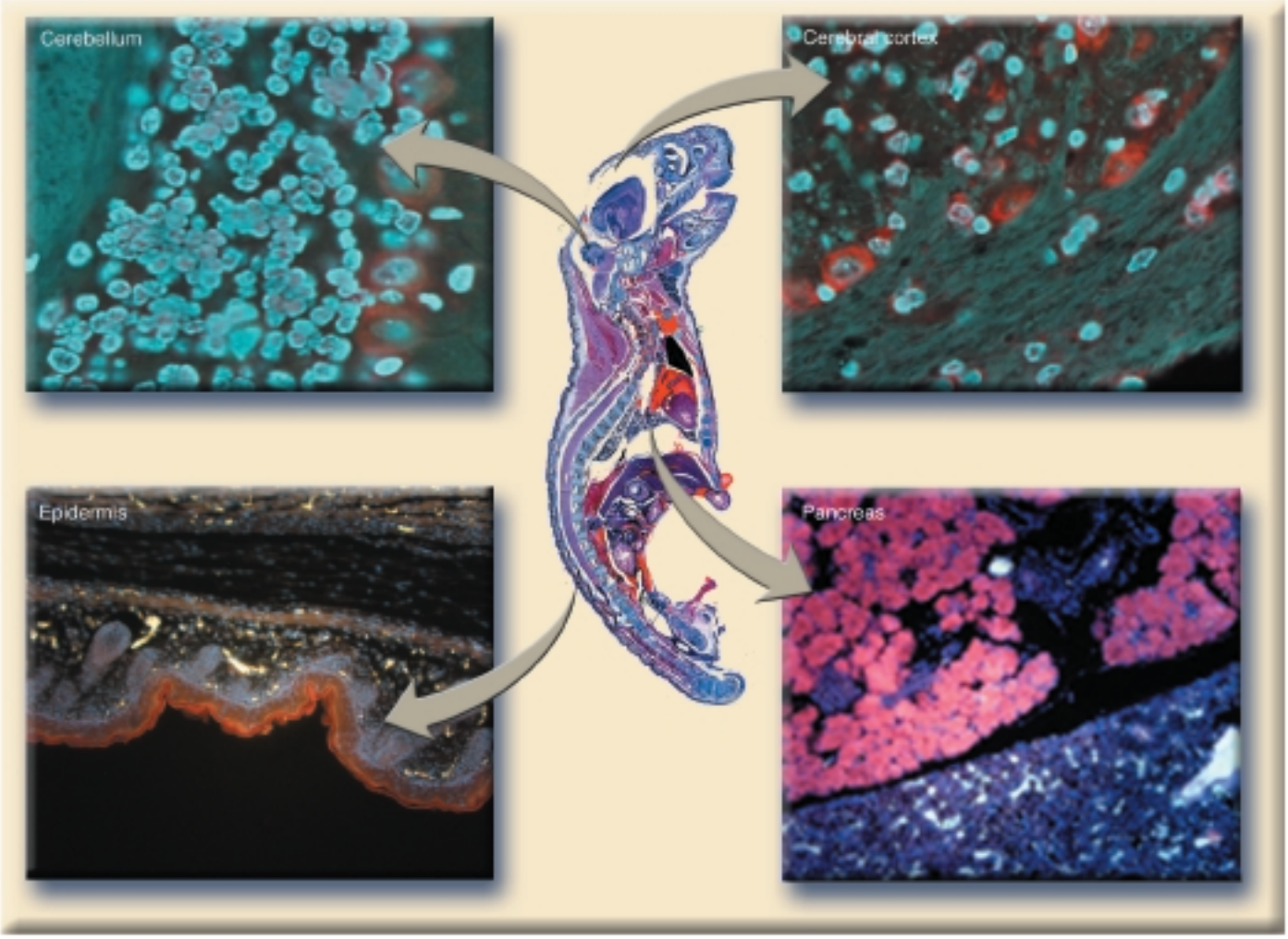


A pip plot comparing the human APLP1 gene with its mouse counterpart. A high degree of similarity is shown between regions of human and mouse exons—the protein-coding DNA sequence of a gene. The exons are indicated by the black boxes at the top of the plot that are numbered from 1 to 16. The matches between human and mouse exons are marked by dots or lines. They indicate similarity generally over 75 percent.

on in the heart, brain, and skin cells, we'll see the fluorescence in all those areas, in the exact cells that are activated. Finding out the exact cell type is important, because two cells that carry out the same function—say, secretion—may be more similar to each other than two different adjacent cells in the same tissue. For example, when we want to know what a gene does, it is much more important to know that the gene is expressed in a Purkinje cell, which helps regulate movement, than to know it's expressed somewhere in the thousands of different cell types that make up the brain."

One gene that was examined in this manner turns out to be activated in only a small section of mouse sequence from a family line extensively studied by Stubbs, where the mice are prone to both deafness and stomach cancer. "What we found out about this gene through section in situ hybridization makes perfect sense to us," says Stubbs. "The gene expresses a protein that protects the epithelial cells lining the insides of body cavities, for example, the stomach. The cells lining the inner ear are also delicate and may require the same kind of protection. We theorize that this same protein performs a similar protective function inside the ear. We haven't proved it, but we think that's why our mice are deaf and have stomach cancer."

Because a single specimen provides 1,000 tissue slices, it can be used to test many genes. Stubbs and her team can create a probe of any gene found on the sequence—whether its purpose is known or unknown—and pinpoint



Thin slices of mouse tissue are placed on a slide, and a gene probe—a specific gene sequence with a fluorescent dye—is added. When the tissue is observed under a microscope, the fluorescent probe can be seen binding to and highlighting the unique gene sequence being studied. Here, this highlighting is shown for gene sequences in the cerebellum, cerebral cortex, epidermis, and pancreas.

where it is expressed, down to type and location of a single cell, in the specimen.

Elsewhere in the comparative genomics community, researchers are focusing on using microarrays to rapidly discover what genes express in tissues or tissue regions and to examine many genes in parallel. However, microarrays do not provide information about the type and location of a cell within a tissue that is expressing a gene or what that cell's context is in the living tissue. "With tissue-section-based techniques, we see exactly where a gene is turned on and can correlate it with the knowledge that pathologists have about what that particular cell does. We can also begin to correlate the state of the gene—its expression patterns in specific types of cells—with its regulatory sequences. This is completely unknown territory."

Stubbs and her team are working to industrialize this process. (See the box on p. 18.) With so many genes to look at, they need to generate a huge amount of information about gene expression to make generalizations about the genes and their regulatory controls. The team is now going through the sequence, looking and testing for candidate versions of these control sequences. "We're beginning to develop some testing techniques that will help us here. Ultimately, we want to go through the chromosome, find these control elements, prove that they are control elements, and then try to correlate expression patterns among them."

New Frontiers Within

If nothing else, all the questions and possibilities just show that, even with the progress scientists have made in piecing together the story of life embedded in the DNA code, complete understanding still eludes them. "The human sequence means absolutely nothing when viewed by itself," notes Stubbs. "We can do very little with it. We can find some of the genes from the expressed sequences we already know about. But we can't read it. We can't figure out where the important sequences are; we miss a lot of the genes; we miss all of the control sequences. What comparative sequence analysis allows us to do is to 'light up' the functional parts of the sequence. If a piece of DNA has an important function, evolution won't let it change. That's the important message in all this. But if we can't find the piece that is doing something important, we won't get very far in our understanding."

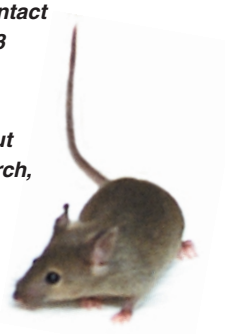
Why does this matter? Consider the gene tied to muscular dystrophy. When the gene is removed from the mouse, the mouse survives. It's a bit uncoordinated, Stubbs says, but it can move around, get on with its life, and reproduce. But when the gene is missing or malfunctioning in humans, the result is a disease of devastating proportions. "Obviously, this gene is much more important to humans than to mice," says Stubbs. "And looking at the differences between the genes and the proteins and how they are regulated in mouse and human will help us understand what part of the human protein is most important. Now we'll be able to do the same sort of analysis for an entire chromosome, thanks to the mouse."

—Ann Parker

Key Words: chromosome 19, comparative genetics, DNA, Human Genome Project (HGP), gene expression, Joint Genome Institute (JGI), mouse genome, PEG3, sequencing, section in situ hybridization.

For further information contact Lisa Stubbs (925) 422-8473 (stubbs5@llnl.gov).

For more information about DOE-funded genetic research, see these Web sites:
www-bio.llnl.gov/genome/
www.jgi.doe.gov/
www.ornl.gov/hgmis/



About the Scientist

LISA STUBBS received a B.S. in biology from the University of Puget Sound in Tacoma, Washington, and a Ph.D. in biology from the University of California (UC) at San Diego. She joined Lawrence Livermore in 1997 as a senior staff scientist in the Genomics and Bioengineering Research Division and in the DOE Joint Genome Institute, where Livermore is one of three collaborating national laboratories. She is currently also acting director of the Genomics and Bioengineering Research Division.

Stubbs leads a team studying mouse genomics, specifically the comparative analysis of structure, function, and evolution of genes in related mouse and human chromosome regions. Her research interests include the generation, biological characterization, and molecular mapping of mouse mutants that provide useful models for studying acquired and inherited human diseases. Stubbs has published over 60 papers in professional journals and is on the editorial board of *Mutation Research Genomics*. She serves on several scientific committees, including the UC Davis Cancer Center Internal Advisory Board, the DOE Biology and Environmental Research Advisory Committee, and the National Institutes of Health Human Genome Study Section.